

城市轨道交通客流量预测的统计分析

王 森, 陈彩霞

重庆理工大学数学科学学院, 重庆

收稿日期: 2025年7月21日; 录用日期: 2025年8月11日; 发布日期: 2025年8月25日

摘 要

研究依据2015年到2023年期间重庆市轨道交通的月度客流量数据, 运用自回归积分滑动平均模型 (ARIMA)、长短期记忆网络 (LSTM), 以及ARIMA与LSTM相结合的混合模型, 对2024年的客流量加以预测。

关键词

轨道交通客流量, ARIMA模型, LSTM神经网络, 混合模型

Statistical Analysis of Passenger Flow Forecasting for Urban Rail Transit

Sen Wang, Caixia Chen

School of Mathematical Sciences, Chongqing University of Technology, Chongqing

Received: Jul. 21st, 2025; accepted: Aug. 11th, 2025; published: Aug. 25th, 2025

Abstract

This study utilizes monthly passenger flow data from Chongqing's rail transit system between 2015 and 2023 to forecast passenger volume in 2024. The analysis employs three approaches: An Auto-regressive Integrated Moving Average (ARIMA) model, a Long-Term Short-Term Memory (LSTM) network, and a hybrid model combining both frameworks.

Keywords

Rail Transit Passenger Flow, ARIMA Model, LSTM Neural Network, Hybrid Model



1. 引言

城市轨道交通是城市公共交通的重要组成部分，具有安全性高、运载能力强、低碳环保等特点，已成为支撑经济快速发展和推动城市化进程的重要支柱。重庆轨道交通作为全国轨道交通运输的代表，精确预测轨道交通客流量，是解决山地城市交通管理复杂问题的关键所在[1] [2]。

本文依据 2015 年到 2023 年期间重庆市轨道交通的月度客流量数据，运用自回归积分滑动平均模型 (ARIMA)、长短期记忆网络(LSTM)，以及 ARIMA 与 LSTM 相结合的混合模型，对 2024 年的客流量加以预测。研究先分别构建 ARIMA 模型和 LSTM 模型进行独立的预测分析，接着利用 ARIMA 模型的残差序列构建 LSTM 模型，将两种模型的预测结果综合起来，混合模型结合了 ARIMA 模型在长期趋势预测方面的优势以及 LSTM 模型在适应短期波动方面的强项，提升了预测的精确度和实用性[3]。

2. 模型介绍

2.1. ARIMA 模型

ARIMA 模型全称为自回归移动平均模型(Autoregressive Integrated Moving Average Model，简记 ARIMA)，其中 ARIMA(p,d,q)称为差分自回归移动平均模型，AR 是自回归，p 为自回归项；I 是差分，d 为时间序列成为平稳时所做的差分次数；MA 是移动平均，q 为移动平均项数[4]。

ARIMA(p,d,q)模型的结构为：

$$\begin{cases} \Phi(B)\nabla^d x_t = \Theta(B)\varepsilon_t \\ E(\varepsilon_t) = 0, Var(\varepsilon_t) = \sigma_\varepsilon^2, E(\varepsilon_t \varepsilon_s) = 0, s \neq t \\ E(x_s \varepsilon_t) = 0, \forall s < t \end{cases}$$

2.2. LSTM 模型

LSTM 模型全称为长短期记忆网络(Long Short-Term Memory，简记 LSTM)，是一种特殊的循环神经网络(RNN) [5]。LSTM 专门用于处理和预测时间序列数据中存在的长距离依赖文体。其核心思想是通过门控机制(包括遗忘门、输入门、输出门)和细胞状态(Cell State)，实现对信息保留与遗忘的动态控制，从而捕捉时间序列中的复杂非线性关系。

1) LSTM 单元结构图见图 1：

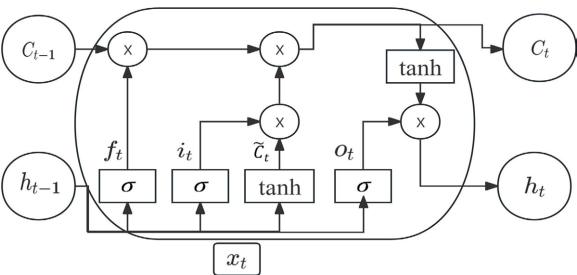


Figure 1. Structure of LSTM unit
图 1. LSTM 单元结构图

2) 遗忘门计算公式

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

3) 输入门计算公式

$$\begin{cases} i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ \tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \end{cases}$$

4) 细胞状态进行更新迭代公式

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$$

5) 输出门计算公式

$$\begin{cases} o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\ h_t = o_t \odot \tanh(C_t) \end{cases}$$

模型介绍部分中涉及各个公式里的参数符号的说明见表 1:

Table 1. Parameter symbol explanation
表 1. 参数符号说明

符号	含义
x_t	当前时刻的输入(如客流量特征)
h_{t-1}	前一时刻的隐藏状态(短期记忆)
C_{t-1}	前一时刻的细胞状态(长期记忆)
W	权重矩阵(可训练参数)
b	偏置项(可训练参数)

2.3. ARIMA-LSTM 混合模型[6]

混合模型(ARIMA-LSTM),其核心想法是:借助 ARIMA 去提取数据当中的线性成分,接着依靠 LSTM 对残差即线性模型没有解释的非线性部分进行建模,最后把两者的预测结果融合起来。

模型结构示意图见图 2:

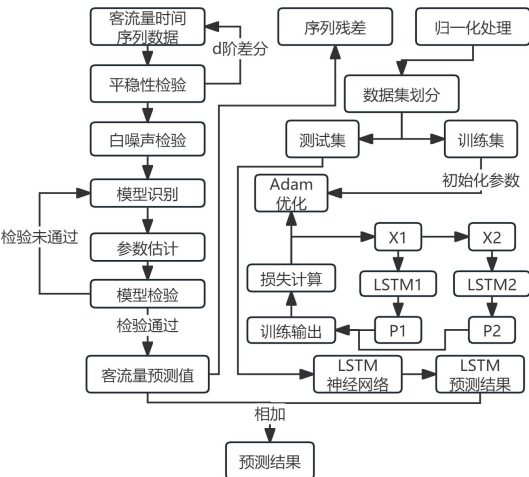


Figure 2. Schematic diagram of model structure
图 2. 模型结构示意图

3. 实证分析

3.1. 数据来源与预处理

1) 数据来源

本研究使用的月度客流量数据覆盖重庆城市轨道交通的全部运营路线(包含地铁、单轨), 时间跨度为2015年1月至2023年12月。数据来源为中国城市轨道交通协会《城市轨道交通统计和分析报告》以及重庆市轨道交通(集团)有限公司发布的《重庆市轨道交通年度运营报告》(2015~2023年)。原始数据见表2所示:

Table 2. Monthly data of rail transit passenger volume in Chongqing from 2015 to 2023 (Unit: 10,000 people)
表 2. 2015~2023 年重庆市轨道交通客运量月度数据(单位: 万人)

时间	客运量	时间	客运量	时间	客运量	时间	客运量	时间	客运量
2015-01	4462.23	2016-11	5920.92	2018-09	7493.14	2020-07	8608.00	2022-05	8774.00
2015-02	3865.86	2016-12	6373.15	2018-10	7724.70	2020-08	8308.00	2022-06	9691.00
2015-03	4911.89	2017-01	5245.61	2018-11	7336.59	2020-09	8755.00	2022-07	10029.00
2015-04	5144.94	2017-02	4538.27	2018-12	7889.67	2020-10	9154.00	2022-08	6602.00
2015-05	5395.58	2017-03	5779.04	2019-01	8140.00	2020-11	9280.00	2022-09	7841.00
2015-06	5149.73	2017-04	6050.01	2019-02	6635.00	2020-12	8943.00	2022-10	7990.00
2015-07	5816.50	2017-05	6345.59	2019-03	8939.00	2021-01	8702.00	2022-11	2252.00
2015-08	6033.73	2017-06	6056.82	2019-04	8724.00	2021-02	6749.00	2022-12	3635.00
2015-09	5515.62	2017-07	6836.22	2019-05	8799.00	2021-03	10072.00	2023-01	-
2015-10	5690.83	2017-08	7091.75	2019-06	8487.00	2021-04	9984.00	2023-02	-
2015-11	5399.98	2017-09	6484.89	2019-07	9226.00	2021-05	10243.00	2023-03	11447.00
2015-12	5813.11	2017-10	6688.76	2019-08	8786.00	2021-06	9668.00	2023-04	11396.00
2016-01	4894.06	2017-11	6348.91	2019-09	8884.00	2021-07	10171.00	2023-05	11377.00
2016-02	4238.74	2017-12	6834.13	2019-10	9303.00	2021-08	7731.00	2023-06	11290.00
2016-03	5392.07	2018-01	6049.11	2019-11	9190.00	2021-09	9233.00	2023-07	11902.00
2016-04	5641.82	2018-02	5256.75	2019-12	9073.00	2021-10	9800.00	2023-08	11269.00
2016-05	5916.30	2018-03	6662.71	2020-01	6957.00	2021-11	7773.00	2023-09	10969.00
2016-06	5650.49	2018-04	6984.94	2020-02	4160.00	2021-12	9593.00	2023-10	12035.00
2016-07	6375.63	2018-05	7326.56	2020-03	3046.00	2022-01	8904.00	2023-11	11497.00
2016-08	6610.69	2018-06	6997.14	2020-04	5772.00	2022-02	7979.00	2023-12	-
2016-09	6047.53	2018-07	7893.98	2020-05	7175.00	2022-03	8823.00		
2016-10	6238.59	2018-08	8184.71	2020-06	7562.00	2022-04	8722.00		

注: 数据包含地铁、单轨等全部制式, 2023 年 1 月、2 月、12 月存在缺失(标记为“-”)。

2) 数据预处理

针对 3 个月份(占比 2.8%)的缺失数据, 采用移动平均法进行插补:
对单个缺失值(如 2023 年 12 月)计算公式为,

$$y_t = \frac{y_{t-1} + y_{t+1}}{2}$$

对连续缺失值(如 2023 年 1~2 月)计算公式为,

$$\hat{y}_{2023-01} = \frac{y_{2022-12} + y_{2023-03}}{2} \times \frac{y_{2022-01}}{2(y_{2021-12} + y_{2022-03})/2}$$

该方法通过相邻时段均值估计缺失值, 保留数据的局部趋势与季节性特征。对于疫情封城以及解封时出现的异常值进行保留, 在后续模型中通过干预变量进行专项处理。

3) 数据特征分析

预处理后数据呈现显著特征其结果见图 3: 1) 长期趋势: 2015~2019 年稳态增长(年均增速 8.5%), 2020 年后受疫情扰动; 2) 季节性: 年度双高峰(7~8 月暑期、10~12 月年末)与春节低谷(1~2 月); 3) 外部事件响应: 2020 年初疫情冲击、2022 年末解封等结构性突变。



Figure 3. Time series of passenger flow after preprocessing
图 3. 预处理后的客流量时序图

3.2. 平稳性检验

Table 3. Results of ADF test on time series
表 3. 时间序列的 ADF 检验结果

延迟阶数	统计量 T 值	P 值
4	-3.3904	0.0659

如表 3 所示, ADF 检验的原假设为“序列非平稳”, 此时 $P = 0.0659 > 0.05$ 没有充分的理由拒绝原假设, 故进行差分处理, 差分结果见图 4 所示, 可看出处理之后的序列是平稳的。

3.3. 模型定阶

设计四个模型, 通过 AIC、BIC 准则选取四个模型相对最优的模型作为轨道交通运量的拟合和预测模型, 四个模型对应的 AIC、BIC 值见表 4 所示。

综合比较摘取的四个模型, $ARIMA(1,1,1) \times (0,0,2)_{12}$ 乘法模型的 AIC、BIC 信息量较小, 是四个模型中相对最优的模型, 因此, 将运用 $ARIMA(1,1,1) \times (0,0,2)_{12}$ 模型展开参数估计、模型检验与预测。

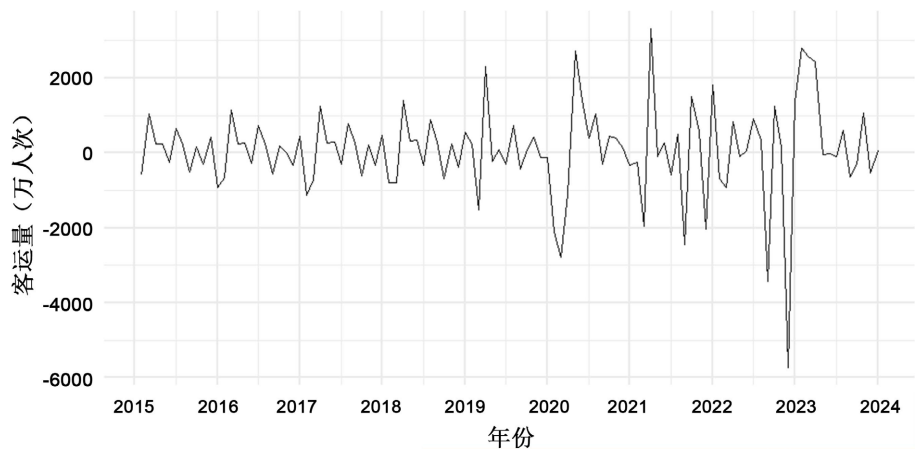


Figure 4. Time series of first-order 12-step difference
图 4. 一阶 12 步差分的时间序列

Table 4. The AIC and BIC values of the four models
表 4. 四个模型对应 AIC、BIC 值

模型形式	AIC 值	BIC 值
ARIMA(1,1,1) × (0,0,2) ₁₂	1826.42	1839.83
ARIMA(1,1,1) × (0,0,1) ₁₂	1830.49	1841.21
ARIMA(0,1,1) × (0,0,1) ₁₂	1833.16	1838.52
ARIMA(0,1,0) × (0,0,1) ₁₂	1834.99	1843.04

其模型的表达式为:

$$(1 - 0.7855B)(1 - B)Y_t = (1 - 0.9518B)(1 + 0.2826B^{12} + 0.2728B^{24})\epsilon_t$$

根据重庆市轨道交通客运量 2015 年 1 月至 2023 年 12 月期间的数据利用模型 ARIMA(1,1,1) × (0,0,2)₁₂ 对轨道交通客运量未来一年数据进行预测, 预测的结果见图 5 所示。

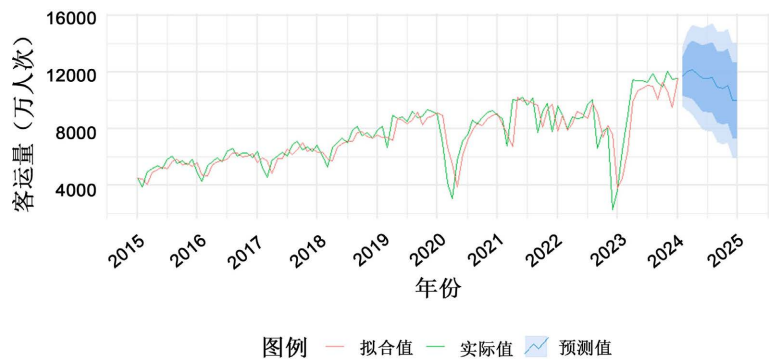


Figure 5. ARIMA model fitting and prediction results
图 5. ARIMA 模型拟合图及预测结果

3.4. 模型优化

1) 干预模型

ARIMA(1,1,1) × (0,0,2)₁₂ 模型所给出的预测结果说明, 在 2018 年至 2023 年这段时间内, 拟合曲线较好地捕捉到客流量所呈现出的季节性波动特征。但模型对于 2020 年 1 月至 2022 年 12 月疫情期间出现的客流异常波动情况, 明显表现出了不适应的状况, 为捕捉 COVID-19 疫情对客流的结构性冲击, 研究采用脉冲干预 ARIMA 模型对城市轨道交通客流量进行建模, 以捕捉突发事件的影响。模型的基础结构为 ARIMA(1,1,1) × (0,0,2)₁₂, 其数学表达式为:

$$\nabla y_t = (1 - 0.6763B)(1 + 0.3912B^{12} + 0.2504B^{24})\varepsilon_t + \sum_i \omega_i I_i(t)$$

其中, 模型引入了四个关键脉冲干预变量见表 5 所示:

Table 5. Parameter estimation of intervention variables
表 5. 干预变量参数估计

变量名称	描述	系数	P 值
初始疫情冲击	2020 年 2 月	-4045.82	P < 0.001
衰减效应	2020 年 3~5 月	-6233.82	P < 0.001
疫情解封	2022 年 11 月	-5108.34	P < 0.001
疫情解封	2022 年 12 月	-4014.79	P < 0.001

实证结果显示, 该模型具有更好的预测性能。模型拟合的 RMSE 为 809.05, 优于原始 ARIMA 模型 (1072.51); MAE 为 742.1731, 优于原始 ARIMA 模型 (565.21); MAPE 为 7.87%, 显著优于原始 ARIMA 模型 (MAPE = 13.94%); AIC 值为 1774.75, 表明模型简洁有效; Ljung-Box 检验 P 值为 0.8797, 证实残差满足白噪声假设。干预修正模型的预测结果见图 6 所示:



Figure 6. Interference model fitting and prediction results
图 6. 干预模型拟合图及预测结果

经过干预修正的 ARIMA 模型, 在各项预测精度指标上呈现出有统计学意义的提升, 虽然干预模型取得了改进, 但其相对改进幅度仍有一定的限制, 这为后续采用混合模型提供了理论依据。

2) LSTM 模型

LSTM 层是整个模型的核心, 包含四个重点考虑对象, 分别是 units (单元数量)、return_sequences 参数、dropout 以及 recurrent_dropout。与 ARIMA 模型不同, 模型中的神经元个数、隐含层层数、时间步长等参数需要通过人为事先设定, 经过多次实验结果, 对比误差不断调节, 得出较好的参数组合。

其数学模型为:

$$\begin{aligned} h_t^1, c_t^1 &= \text{LSTM}_1(x_t, h_{t-1}^1, c_{t-1}^1) \\ h_t^2, c_t^2 &= \text{LSTM}_2(\text{Dropout}(h_t^1), h_{t-1}^2, c_{t-1}^2) \\ y_t &= W \cdot \text{Dropout}(h_t^2) + b \end{aligned}$$

预测结果见图 7 所示:

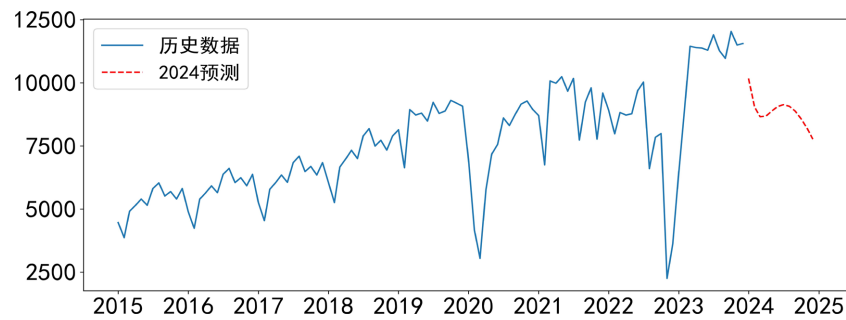


Figure 7. Passenger volume prediction of LSTM model in 2024
图 7. LSTM 模型 2024 年客运量预测

从整体来看, 预测结果呈现出节假日对轨道交通客流量所产生的影响, 同时也体现出疫情对客流量有持续的作用, 使得客流量相对保持平稳, 但低于疫情之前的水平。LSTM 模型在预测上呈现出的局限性, 表现主要源于以下原因: 其一、训练数据仅为 2015~2020 年的月度数据, 应对疫情期间极端事件样本不足, 导致模型对异常值学习不充分; 其二、单一 LSTM 结构对长期趋势和短期波动的协同捕捉能力有限。后续研究将通过采用混合建模框架, 将 LSTM 与 ARIMA 的优势相结合, 加以改进。

3) 混合模型预测

为了保证模型的泛化能力和预测精度, 根据相关学者研究和实践经验, 将残差序列中前 80%数据作为训练集, 后 20%数据作为测试集。预处理后的残差序列见图 8 所示。

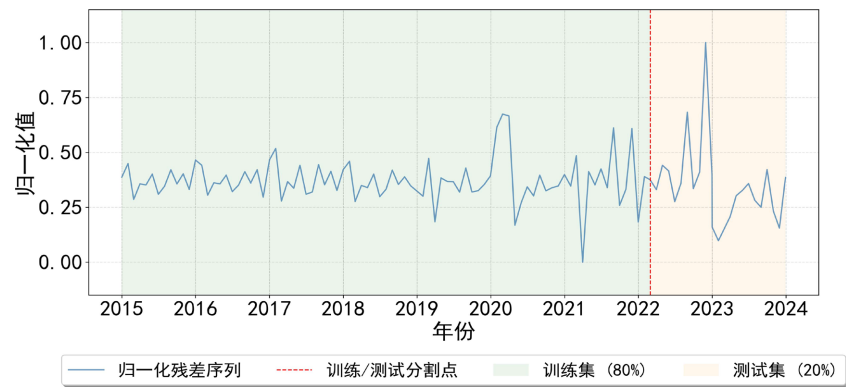


Figure 8. Residual sequence diagram
图 8. 残差序列图

在单一 LSTM 模型中加入特征工程可以提升模型捕捉非线性模式的能力, 故引入上述外部特征: 1) 节假日特征: 标记元旦(1 月)、春节(2 月)、五一(5 月)、国庆(10 月); 2) 疫情冲击特征: 2020~2023 年标

记为 1, 其余为 0, 见图 9 和图 10 所示。



Figure 9. Holiday characteristics and pandemic impact

图 9. 节假日特征及疫情影响

将 ARIMA 模型的预测值与 LSTM 模型对残差序列的预测值相加得到最终预测结果见图 10 所示。



Figure 10. Comparison between actual value and ARIMA-LSTM fitting value

图 10. 实际值与 ARIMA-LSTM 拟合值对比

实证分析显示, ARIMA-LSTM 混合模型在预测性能方面相较于单一模型具有显著优势。该模型的均方根误差(RMSE, 728.04)较基础 ARIMA 模型(1072.51)和 LSTM 模型(1372.51)分别降低了 32.1%和 46.9%; 平均绝对误差(MAE, 515.57)较上述两者分别降低了 30.5%和 44.1%; 平均绝对百分比误差(MAPE, 6.89%)也明显优于基础 ARIMA 模型(13.94%)和 LSTM 模型(19.31%), 优化后的干预 ARIMA 模型(RMSE = 809.05, MAPE = 7.87%)性能仍稍逊于混合模型。此结果验证了混合模型在极端值预测以及常规误差控制方面的双重优越性, 表明其预测精度的稳健特性。研究证实, 融合时序特征提取与非线性学习的混合建模方法能够有效提高预测精度, 为轨道交通客流管理提供了更为可靠的技术支撑。

4. 结论及建议

本文在混合模型构建以及短期预测领域取得了一定成果, 不过, 仍存在诸多可改进与扩展之处, 针对未来研究方向, 给出以下几方面展望: 其一, 在模型改进方面, 当前模型虽已有不错的预测效果, 但仍有潜力依靠引入更多影响因素提升精准度, 可考虑加入天气数据、节假日信息以及周边商业活动具体

情况等多维度数据, 这些数据的融入可使模型更全面地捕捉影响客流变化的复杂因素, 让预测结果更精确可靠。其二, 在实际应用方面, 现有研究主要聚焦于常规情况下的客流预测, 对于特殊情景, 如大型活动举办或极端天气条件下的客流变化, 需深入探讨, 未来研究应着重开发针对这些特殊情景的客流预测方法, 构建更全面灵活的预测系统, 契合不同场景实际需求。

参考文献

- [1] 中国城市轨道交通协会. 城市轨道交通 2023 年度统计和分析报告[R]. 北京: 中国城市轨道交通协会, 2024.
- [2] 重庆市轨道交通(集团)有限公司. 重庆轨道交通线建设运营单位[Z]. 重庆: 重庆市轨道交通(集团)有限公司, 2024.
- [3] 刘海洲, 李荣迪. 重庆市轨道交通车站预测客流与实际客流误差分析[J]. 铁道运输与经济, 2022, 44(4): 119-123.
- [4] Ying, S.L., Shuo, L., Ye, T., *et al.* (2021) Research on Forecast of Rail Traffic Flow Based on ARIMA Model. *Journal of Physics: Conference Series*, **1792**, Article 012065. <https://doi.org/10.1088/1742-6596/1792/1/012065>
- [5] David, D.C. and Weiya, C. (2022) Forecasting Daily and Weekly Passenger Demand for Urban Rail Transit Stations Based on a Time Series Model Approach. *Forecasting*, **4**, 904-924. <https://doi.org/10.3390/forecast4040049>
- [6] Li, W., Sui, L.Y., Zhou, M., *et al.* (2021) Short-Term Passenger Flow Forecast for Urban Rail Transit Based on Multi-source Data. *EURASIP Journal on Wireless Communications and Networking*, **2021**, Article No. 9. <https://doi.org/10.1186/s13638-020-01881-4>